项目反应理论观察分数核等值的影响因素

王少杰　张敏强*　黄菲菲　黄丽芳　袁琪婷

（华南师范大学心理学院，广州，510631）

摘要 探究带宽选择方法、样本量、题目数量、等值设计、数据模拟方式对项目反应理论观察分数核等值的影响。通过两种数据模拟方式，获得研究数据，并计算局部与全域评价指标。研究发现，在随机组设计中，带宽选择方法表现相似；考生样本量和题目数量影响甚微。在非等组设计中，惩罚法与 Silverman 经验准则表现优异；增加题目量可降低百分相对误差和随机误差；增加样本量导致百分相对误差变大，随机误差减小。数据模拟方式可影响等值评价。未来应重点关注等值系统评估。

关键词 IRT 观察分数核等值；带宽选择方法；等值设计；数据模拟方式

1 问题提出

核等值（Kernel Equating，KE）是一种测验等值方法体系，它基于近似传统等百分位等值（Equipercentile Equating，EE），并将线性等值作为特例（von Davier et al., 2004）。研究流程共包含五步：（1）预平滑，即采用对数线性模型拟合观察分数。（2）估计分数概率，即通过设计函数，将拟合的样本分数概率转化为总体分数概率。（3）连续化，即将离散累积分布连续化。（4）等值，即采用 EE 计算结果。（5）计算等值标准误等指标，即评估等值结果（罗莲，2008）。KE 采用预平滑和连续化，可降低因样本量较少造成的随机误差（Jiang et al., 2012; von Davier & Chen, 2013）。从等值设计到等值评价，均在一系列特有且相互联系的框架中完成；同时可对各环节单独分析，调整参数，得到与其他方法相似的结果，极具包容与扩展性（王少杰等，2020）。综上，KE 相较于经典测量理论（Classical Test Theory，CTT）等值更为精确、稳定，具有较大发展和应用前景。

但由于其采用 EE 计算结果，不能从更微观角度，将考生与题目参数同时建模，而这恰为项目反应理论（Item Response Theory，IRT）等值擅长之处。在分数层面，IRT 等值方法包含观察分数等值（IRT Observed Score Equating，IRTOSE）与真分数等值（IRT True Score Equating，IRTTSE）。在将参数置于同一量尺后，前者通过 IRT 模型计算待等两测验正确作答分数分布，最后采用 EE 获得分数对应关系；而后者认为两测验中相同能力对应的真分数即为等值分数（Kolen & Brennan, 2014）。KE 与 IRT 等值方法各有优劣，在特定条件下，前者优于后者（De Ayala et al., 2018; Leôncio & Wiberg, 2017; Wang et al., 2020）。那么，是否可

* 通讯作者：张敏强。E-mail: 2640726401@qq.com。

将 KE 与 IRT 等值方法结合，使新方法既有前者的连续化思想，又包含后者的优异特性呢？

2 IRT 观察分数核等值及其相关概念

2.1 IRT 观察分数核等值

Andersson 等（2013）最早将 KE 与 IRTOSE 结合，在专门开展 KE 分析的 kequate 软件包中提出 IRT 观察分数核等值（IRT observed score Kernel Equating，IRTKE）的概念。因非等组锚测验设计（Non-Equivalent groups with Anchor Test design，NEAT）较为常用，且本研究亦基于此，故仅介绍 NEAT 下的 IRTKE。其他设计可参考 Andersson 和 Wiberg（2017）和 Sansivieri 等（2017）。

假设有测验 X、Y 与锚测验 A，用r_j、s_k、$t_{P,l}$、$t_{Q,l}$分别代表来自群体 P 和 Q 的考生分别取得分数x_j、y_k和a_l的概率。首先，采用 IRT 模型（本研究为二参数逻辑斯蒂克模型，2-Parameter Logistic Model，2PLM）拟合数据，得到考生和题目参数。第二步，依据$r_j \approx \sum_{r=1}^{R} r_j(t_r)W(t_r)$计算得分概率，$t_r$为第 r 个积分节点的能力水平，$r_j(t_r)$为根据 Lord 和 Wingersky （1984）迭代算法求出的分数概率，$W(t_r)$为积分节点t_r的权重。s_k、$t_{P,l}$与$t_{Q,l}$算法类似。第三步，连续化。将离散变量 $\boldsymbol{X}$ 连续化后，可得到$\boldsymbol{X}(h_X)$的累积分布函数$F_{h_X}(x) = \sum_j r_j \Phi\left(R_{jX}(x)\right)$，其中$\Phi(z)$为标准正态分布的累积分布函数，$R_{jX}(x) = \frac{x - a_X x_j - (1-a_X)\mu_X}{a_X h_X}$，$h_X$为带宽。$\boldsymbol{Y}$ 与 $\boldsymbol{A}$ 的连续化与之类似。第四步，等值。依据 EE，$\hat{e}_{Y(CE)}(x) = \hat{G}^{-1}_{Qh_{YQ}}(\hat{H}_{Qh_{AQ}}(\hat{H}^{-1}_{Ph_{AP}}(\hat{F}_{Ph_{XP}}(x))))$，其中，$\hat{G}^{-1}_{Qh_{YQ}}(\cdot)$、$\hat{H}_{Qh_{AQ}}(\cdot)$、$\hat{H}^{-1}_{Ph_{AP}}(\cdot)$、$\hat{F}_{Ph_{XP}}(\cdot)$分别为测验 Y 和 A 在 Q 上的分数分布，以及测验 A 和 X 在 P 上的分数分布。第五步，计算等值标准误（Standard Error of Equating，SEE）和等值差异标准误（Standard Error of Equating Difference，SEED）。

Andersson（2016）推导出多级计分 IRTKE 的渐近标准误，为等值评估提供可靠指标。Wiberg（2016a）提出局部线性 IRTKE，用 IRT 模型拟合数据，再利用作答反应概率求得线性等值中的总体参数进行等值。Wiberg（2016b）比较了传统方法与 IRTKE 的表现，但未得出明确结论。Andersson 和 Wiberg（2017）系统介绍了基于 IRTKE 的链等值和后分层等值，发现其等值标准误和偏差均较小。Wang 等人（2020）模拟操纵了样本量、测验长度、数据模拟方式与参照等值，比较 EE、KE、IRTOSE 与 IRTKE，发现在随机组设计（Equivalent Groups design，EG）下 IRTKE 表现最优，而在 NEAT 下，IRTOSE 最优。但尚未有学者专门探讨 IRTKE 的影响因素。

2.2 连续化与带宽选择方法

连续化是 KE 的关键。将 $\boldsymbol{X}$ 与连续变量 $\boldsymbol{V}$ 加和，得到$\boldsymbol{X}(h_X) = a_X(\boldsymbol{X} + h_X\boldsymbol{V}) +$

$(1-a_X)\mu_X$，其中$a_X^2=\frac{\sigma_X^2}{\sigma_X^2+h_X^2}$，$h_X$为带宽，$\mu_X$与$\sigma_X^2$为 $\boldsymbol{X}$ 的平均数与方差。$\boldsymbol{Y}(h_Y)$转换同理。可见，h_X控制着分数连续化程度。最常用的带宽选择方法为惩罚法（Penalty Method），$PEN(h_X)=\sum_j\left(r_j-f_{h_X}(x_j)\right)^2+K\cdot\sum_j A_j(1-B_j)$，其中$\boldsymbol{X}(h_X)$的概率密度函数$f_{h_X}(x)=\sum_j r_j\phi\left(R_{jX}(x)\right)\frac{1}{a_X h_X}$，$\phi(\cdot)$为标准正态分布的概率密度函数；K 为常数，当在$x_j$稍偏左位置$f'_{h_X}(x)<0$时，$A_j=1$，当在$x_j$稍偏右位置$f'_{h_X}(x)>0$时，$B_j=0$。

Silverman 经验准则（Silverman's Rule of Thumb method）通过使渐近平均积分平方误差最小化从而求取带宽（Andersson & von Davier, 2014）。为减小异常值的影响，避免过度平滑；同时考虑a_X，得到$h_X=\frac{9\sigma_X}{\sqrt{100n_X^{\frac{2}{5}}-81}}$，其中，$\sigma_X$含义同上，$n_X$为参加测验 X 的考生人数。

Häggström 和 Wiberg（2014）将重复平滑法（Double Smoothing Method）应用于带宽选择。首先，以最小分数单位的一半进行连续化，得到$f_{gX}(x)=\sum_{j=1}^{J}r_j\phi\left(R_{jgX}(x)\right)\frac{1}{a_{gX}h_{gX}}$，其中，$R_{jgX}(x)=\frac{x-a_{gX}x_j-(1-a_{gX})\mu_X}{a_{gX}h_{gX}}$，$a_{gX}^2=\frac{\sigma_X^2}{\sigma_X^2+h_{gX}^2}$，其余指标含义与上文一致或相似。其次，利用$f_{gX}(x)$，再次连续化得到$f_{h_X}(x)=\sum_{j=1}^{J}f_{gX}(x_j)\phi\left(R_{jX}(x)\right)\frac{1}{a_X h_X}$。最后，求使重复平滑函数最小对应的带宽$DS(h_X)=\sum_{l=1}^{2J-1}\left(\hat{r}_l-f_{h_X}(x_l)\right)^2$，其中$\hat{r}_l=\begin{cases}r_{\frac{l+1}{2}}, & l\text{为奇数}\\ f_{h_X}(x_l), & l\text{为偶数}\end{cases}$。

虽有研究表明，三种带宽选择方法均较为优异（Andersson & von Davier, 2014; Häggström & Wiberg, 2014），但还未有研究探讨其对 IRTKE 的影响。

2.3 数据模拟方式

在等值模拟研究中，一般借助蒙特卡洛方法，从特定的先验分布中抽取参数值，通过 IRT 模型，计算被试作答反应矩阵。该方式便捷、易理解和接受。但研究结论可能偏向 IRT 等值方法（De Ayala et al., 2018； Norman Dvorak, 2009）。如何降低数据模拟方式可能导致的偏差呢？构造伪测验及伪群组法（Pseudo-Test Forms and Pseudo-Groups）最早由 Petersen 等（1982）提出，通过抽取实测数据，构建虚拟测验与考生；尽量保证模拟的现实性而又不失结论的一般性。虽有不少学者将其应用于等值比较（例如，Hagge & Kolen, 2012; Kim & Lu, 2018; Powers & Kolen, 2012），但均未涉及 IRTKE。验证该方法的可靠性，可为相关研究提供新的切入点。

3 研究设计

3.1 研究目的

采用不同数据模拟方式和等值设计，探究带宽选择方法、样本量与题目数量对 IRTKE

的影响。

3.2 纳入的影响因素

3.2.1 带宽选择方法

惩罚法、Silverman 经验准则、重复平滑法为三种常用的带宽选择方法，对它们相互比较的同时，亦可提高结论外部效度。

3.2.2 考生样本量

IRT 参数估计对样本量要求较高，一般取 500 以上（Hambleton & Jones, 1993）。结合相关研究（De Ayala et al., 2018; Kim, 2014; Liang & von Davier, 2014），共设定三个样本量水平：1000 人（小样本）、2000 人（中等样本）、5000 人（大样本）*。

3.2.3 题目数量

根据 De Ayala 等人（2018）、Kim（2014）、Liang 和 von Davier（2014）及国内考试的试卷构成，并结合抽样数据源情况（见表 1），设定题目量水平：30 与 45，分别代表较短和中等长度测验。NEAT 锚题比例设定为 30%，即 9 题与 14 题。

3.2.4 等值设计

EG 与 NEAT 涵盖常用等值设计的处理思想，故在二者情况下探究 IRTKE 的表现。NEAT 中具体采用的 EE 方法为链等值（Chained Equating）。

3.2.5 数据模拟方式

构造法直接在真实数据中抽样以得到满足特定要求的测验与考生样本。IRT 法需在特定的先验分布中抽取参数值，从而计算作答矩阵。综合考量二者结果，可提高结论的普适性。

3.3 数据与工具

数据源为某大型语言测试 Form 1 和 Form 2（同一批考生；González & Wiberg, 2017），各包含 80 道题（二级计分）。两次测试间隔 6 个月，基础情况见表 1。采用 2PLM 拟合 Form 1 数据，得到题目参数信息见表 2。

表 1 语言测试

	Form 1	Form 2		Form 1	Form 2
样本量	8000	8000	标准差	12.66	12.59
题目数量	80	80	偏度	0.12	0.04
最低分（理论值）	9（0）	11（0）	峰度	-0.65	-0.65
最高分（理论值）	79（80）	78（80）	信度	0.90	0.90
平均值	43.33	44.24	相关	0.86	

表 2 Form1 题目参数

* 本研究采用 EG 与 NEAT，单次等值需两个考生样本，故此处为其样本量之和。

	平均数	标准差	最小值	最大值
区分度	0.78	0.28	0.15	1.47
难度	0.25	0.74	-1.56	2.13

所有分析采用 R（R Core Team, 2017）调用软件包 kequate（Andersson et al., 2013）、mirt（Chalmers, 2012）、equateIRT（Battauz, 2015）完成。

3.4 模拟流程

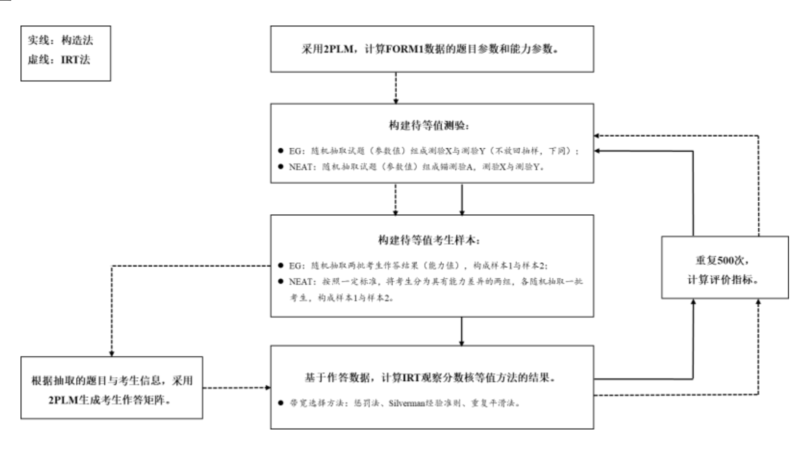


图 1　模拟流程

以 NEAT 为例，详细介绍两种数据模拟方式的研究过程（EG 较易，不赘述）。

3.4.1 构造法

（1）在 Form1 中随机抽取试题及全部考生对应作答结果，构造锚测验 A、测验 X 与 Y。（2）按 Form2 成绩是否高于平均分，将考生分成高、低分数组，分别随机抽取考生及其作答结果，构成具有能力差异的两样本。（3）采用不同带宽选择方法，计算 IRTKE。（4）重复上述过程 500 次，计算评价指标。

3.4.2 IRT 法

（1）采用 2PLM 拟合 Form1 数据，并计算题目与被试参数。（2）随机抽取试题及对应试题参数估计值，构造锚测验 A、测验 X 与 Y。（3）根据能力均值将考生划分成高、低能力组，分别随机抽取能力值，构成具有能力差异的两样本。（4）通过 2PLM，计算考生正确作答概率，并将其与从 $U(0,1)$ 抽取的随机数比较，获得作答矩阵。（5）采用不同带宽选择方法，计算 IRTKE。（6）重复上述过程 500 次，计算评价指标。

3.5 评价指标

3.5.1 局部指标

局部指标有百分相对误差（Percent Relative Error，PRE）与 SEE，反映等值方法在单个分数点或原点矩的表现。

（1）PRE 度量$e_Y(\boldsymbol{X})$与$\boldsymbol{Y}$的分布差异，p阶 PRE 为$PRE(p) = 100 \times \frac{\mu_p(e_Y(\boldsymbol{X}))-\mu_p(\boldsymbol{Y})}{\mu_p(\boldsymbol{Y})}$，其中$\mu_p(\boldsymbol{Y}) = \sum_k (y_k)^p s_k$，$\mu_p(e_Y(\boldsymbol{X})) = \sum_k \left(e_Y(x_j)\right)^p r_j$。

采用链等值可分别得到将测验 X 等值到 A 与将锚测验 A 等值到 Y 的 PRE。将二者取绝对值后加和，并求 500 次的平均值。

（2）SEE 代表随机误差，$SEE_Y(x) = \parallel \boldsymbol{J}_{e_Y}\boldsymbol{J}_{DF}\boldsymbol{C} \parallel$，其中，$\boldsymbol{J}_{e_Y}$为 KE 函数关于 $\boldsymbol{r}$ 与 $\boldsymbol{s}$ 的雅可比矩阵；$\boldsymbol{J}_{DF}$为设计函数关于 $\boldsymbol{r}$ 与 $\boldsymbol{s}$ 的雅可比矩阵；$\boldsymbol{C}$为估计分数概率阶段得到的特殊矩阵，且$\boldsymbol{\Sigma}_{r,s} = Cov\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{s}\end{pmatrix} = \boldsymbol{C}\boldsymbol{C}^t$；$\parallel \boldsymbol{v} \parallel = \sqrt{\sum_j v_j^2}$。最后，将 500 批 SEE 求平均。

3.5.2 全域指标

全域指标刻画等值方法在分数区间或所有原点矩的表现，包含平均 PRE（Averaged PRE，APRE）与平均 SEE（Averaged SEE，ASEE）。$APRE = \sum_{p=1}^{10} PRE(p)$，$ASEE = \sum_i w_i SEE_Y(x_i)$，其中，$w_i = \frac{N_i}{N_T}$，$N_i$为分数$x_i$的人数，$N_T$为总人数。

4 结果

4.1 抽样概况

分别计算各条件组合得到的 500 批测验 X 分数的描述统计量，详见表 3 和表 4。

表 3　模拟分数情况（EG）

模拟方式	样本量-题目量	平均数	标准差	最小值	最大值	偏度	峰度
构造法	1000-30	16.29	5.18	0	30	0.06	-0.59
	2000-30	16.28	5.19	0	30	0.06	-0.60
	5000-30	16.28	5.19	0	30	0.06	-0.60
	1000-45	24.43	7.45	1	45	0.08	-0.63
	2000-45	24.42	7.45	1	45	0.08	-0.63
	5000-45	24.42	7.44	1	45	0.08	-0.63
IRT 法	1000-30	13.35	5.58	0	30	0.45	-0.35
	2000-30	13.36	5.58	0	30	0.45	-0.35
	5000-30	13.35	5.57	0	30	0.45	-0.35
	1000-45	20.04	8.06	0	45	0.49	-0.33
	2000-45	20.05	8.06	0	45	0.49	-0.33
	5000-45	20.05	8.06	0	45	0.49	-0.33

表 4　模拟分数情况（NEAT）

模拟方式	样本量-题目量	平均数	标准差	最小值	最大值	偏度	峰度
构造法	1000-30	19.78	4.04	0	30	-0.21	-0.11
	2000-30	19.78	4.03	0	30	-0.21	-0.10
	5000-30	19.78	4.03	0	30	-0.22	-0.11
	1000-45	29.67	5.63	3	45	-0.19	-0.06
	2000-45	29.67	5.63	2	45	-0.19	-0.07

	5000-45	29.67	5.64	2	45	-0.19	-0.07
IRT 法	1000-30	17.77	4.40	2	30	0.26	-0.40
	2000-30	17.78	4.40	2	30	0.27	-0.39
	5000-30	17.78	4.40	2	30	0.27	-0.39
	1000-45	26.66	6.18	7	45	0.38	-0.36
	2000-45	26.66	6.19	8	45	0.38	-0.36
	5000-45	26.66	6.18	6	45	0.38	-0.36

4.2 EG

4.2.1 局部表现

据图 2，随原点矩阶数升高，PRE 均不同程度上扬，等值前后分数分布形态差异逐渐增大。除高阶原点矩和个别情况（图 2 左下中“sil-1000-30”）外，带宽参数选择方法对分数转换的影响基本无差异。扩大样本量与题目量对降低 PRE 指标有积极作用，其中后者尤为明显（对比图 2 左上与右上、左下与右下）。当采用 IRT 法且 30 道题时，各 PRE 曲线较分散，数据模拟方式与题目数量间存在交互作用（对比图 2 左上与左下、右上与右下）。

从 SEE 角度（图 3），与 PRE 类似，带宽参数选择方法间 SEE 基本无差异，扩大样本量可减小随机误差。但题目量增加反而导致 SEE 稍有扩大。此外，采用构造法得到的 SEE“左高右低”，而 IRT 法却为“左低右高”。

4.2.2 全域表现

据表 5，在 EG 中，APRE 较小（除 0.27 外），带宽参数选择方法间 APRE 无明显差异。样本量与题目数量对其影响与前述一致。采用构造法的 APRE 较 IRT 法小，但差异甚微。

各带宽选择方法间的 ASEE/SD 相同，且随样本量增加和题目量减少，有降低趋势，但幅度较小。构造法与 IRT 法所得 ASEE 基本无差异。

表 5　全域指标 APRE 与 ASEE（EG）

样本量-题目量	APRE						ASEE/SD[*]					
	构造法			IRT 法			构造法			IRT 法		
	Pen	Sil	Dou	Pen	Sil	Dou	Pen	Sil	Dou	Pen	Sil	Dou
1000-30	0.04	0.04	0.04	0.06	0.27	0.06	0.03	0.03	0.03	0.03	0.03	0.03
2000-30	0.02	0.02	0.02	0.04	0.04	0.04	0.02	0.02	0.02	0.02	0.02	0.02
5000-30	0.01	0.01	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
1000-45	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.04	0.05	0.05	0.05
2000-45	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03
5000-45	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02

注：Pen、Sil 与 Dou 分别代表惩罚法、Silverman 经验准则、重复平滑法。

[*] 为便于结果间比较，采用原始测验分数的标准差对 ASEE 标准化处理。表 6 同理。

4.3 NEAT

4.3.1 局部表现

从 PRE 角度，据图 4，随原点矩阶数升高，PRE 呈现不同程度上扬。惩罚法与 Silverman 经验准则结果重合，对分数转换影响较小；且当题目数量较少时，该优势较突出（对比图 4 左上与右上、左下与右下）。与 EG 相似，题目量增加，PRE 减小；但样本量却几乎不影响 PRE。采用 IRT 法且题目量为 30 时，重复平滑法的 PRE 受考生人数影响较大。同时，基于构造法的 PRE 值较 IRT 法对应值小。

从 SEE 角度，据图 5，采用构造法时，带宽参数选择方法表现基本一致，仅在高、低分处存在略微差异。采用 IRT 法时，带宽参数选择方法仅在中间偏低分数处相似；在高、低分数附近，Silverman 经验准则逊于其他方法；而在中间偏高分数处，优于其他方法。扩大样本量和题目量可减小随机误差。同时，构造法的 SEE 在数值和形态波动上均较 IRT 法小。

4.3.2 全域表现

参照表 6，在 NEAT 中，APRE 值较大（最大 3.52）。与局部表现一致，惩罚法和 Silverman 经验准则的 APRE 较重复平滑法小。样本量增加，APRE 变大，但增量不明显；而题目量增加对降低 APRE 有显著作用。构造法的 APRE 较 IRT 法对应值小。

ASEE 指标在带宽选择方法间不存在明显差异，且其随样本量与题目量的增加，均呈减小趋势。构造法的 ASEE 远小于 IRT 法对应值。

表 6　全域指标 APRE 与 ASEE（NEAT）

样本量-题目量	APRE						ASEE/SD					
	构造法			IRT 法			构造法			IRT 法		
	Pen	Sil	Dou	Pen	Sil	Dou	Pen	Sil	Dou	Pen	Sil	Dou
1000-30	2.07	2.07	2.75	2.29	2.29	3.17	0.68	0.69	0.69	2.79	2.63	2.82
2000-30	2.09	2.09	2.72	2.31	2.31	**3.52**	0.38	0.38	0.38	0.94	0.88	0.96
5000-30	2.11	2.11	2.73	2.33	2.33	3.33	0.28	0.29	0.28	0.43	0.41	0.44
1000-45	0.96	0.96	1.22	1.19	1.19	1.62	0.13	0.13	0.13	2.69	2.56	2.71
2000-45	0.97	0.97	1.23	1.21	1.21	1.64	0.05	0.05	0.05	0.38	0.36	0.38
5000-45	0.98	0.98	1.24	1.22	1.22	1.66	0.03	0.03	0.03	0.04	0.04	0.04

注：Pen、Sil 与 Dou 分别代表惩罚法、Silverman 经验准则、重复平滑法。

5 讨论

5.1 带宽选择方法与等值设计

带宽选择方法对等值的影响，因等值设计而异。在 EG 中，带宽选择方法间无较大差异（Häggström & Wiberg, 2014）；但在 NEAT 中，惩罚法与 Silverman 经验准则的表现优于重

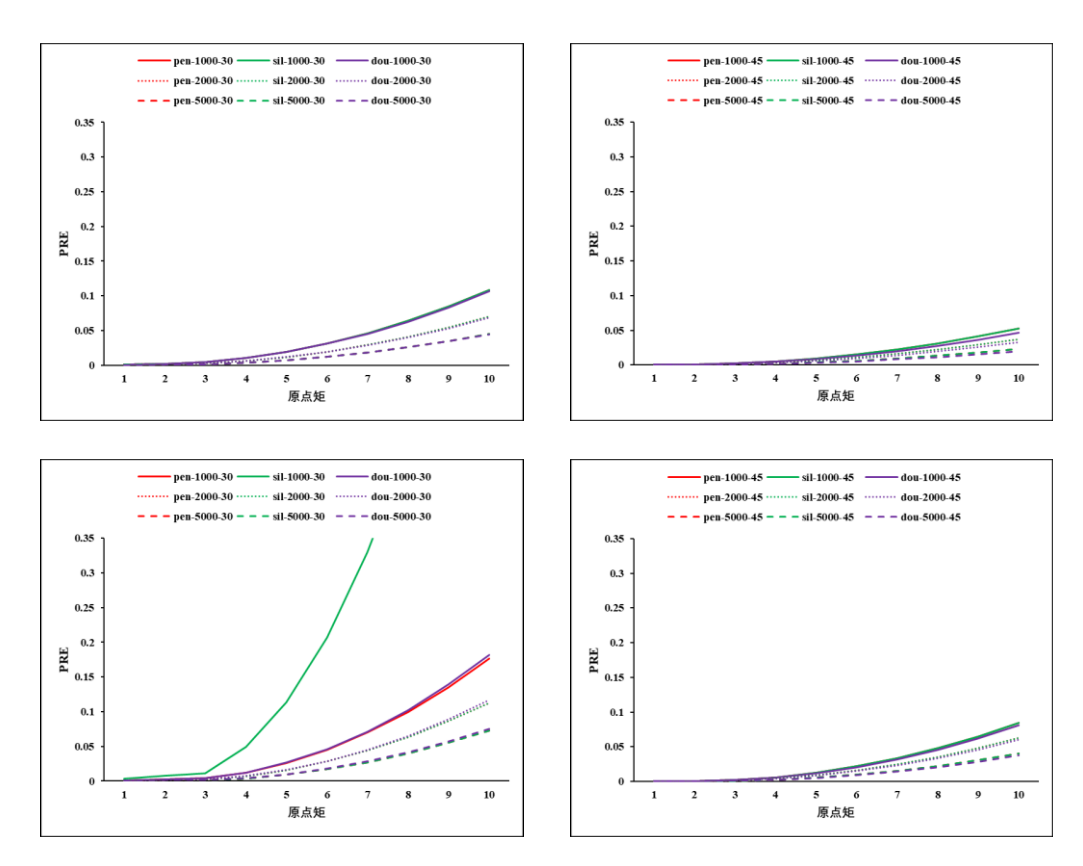


图 2　PRE（EG）

注：上图与下图分别为构造法和 IRT 法；Pen、Sil 与 Dou 分别代表惩罚法、Silverman 经验准则、重复平滑法；1000、2000、5000 与 30、35 分别代表样本量和题目量；除左下外，其他各图中三种带宽选择方法结果基本重合。

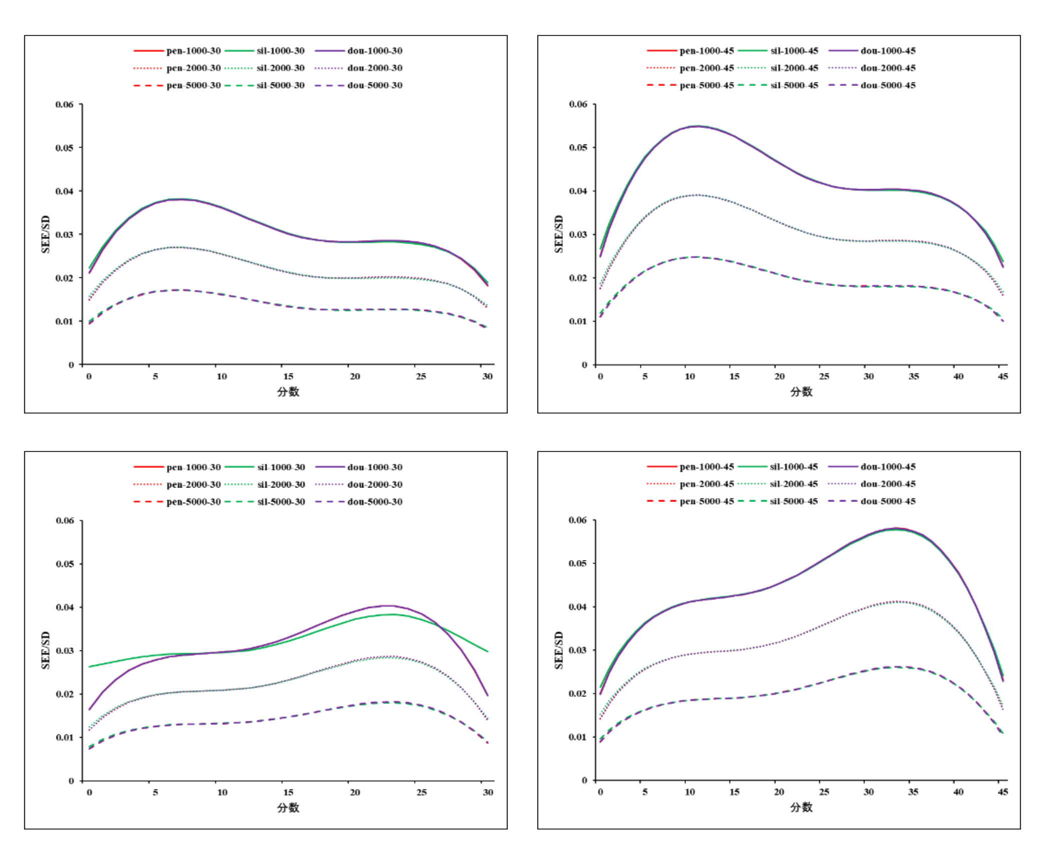


图 3　SEE（EG）

注：上图与下图分别为构造法和 IRT 法；Pen、Sil 与 Dou 分别代表惩罚法、Silverman 经验准则、重复平滑法；1000、2000、5000 与 30、35 分别代表样本量和题目量；除左下外，其他各图中三种带宽选择方法结果基本重合；纵轴将所有结果置于标准差单位量尺。

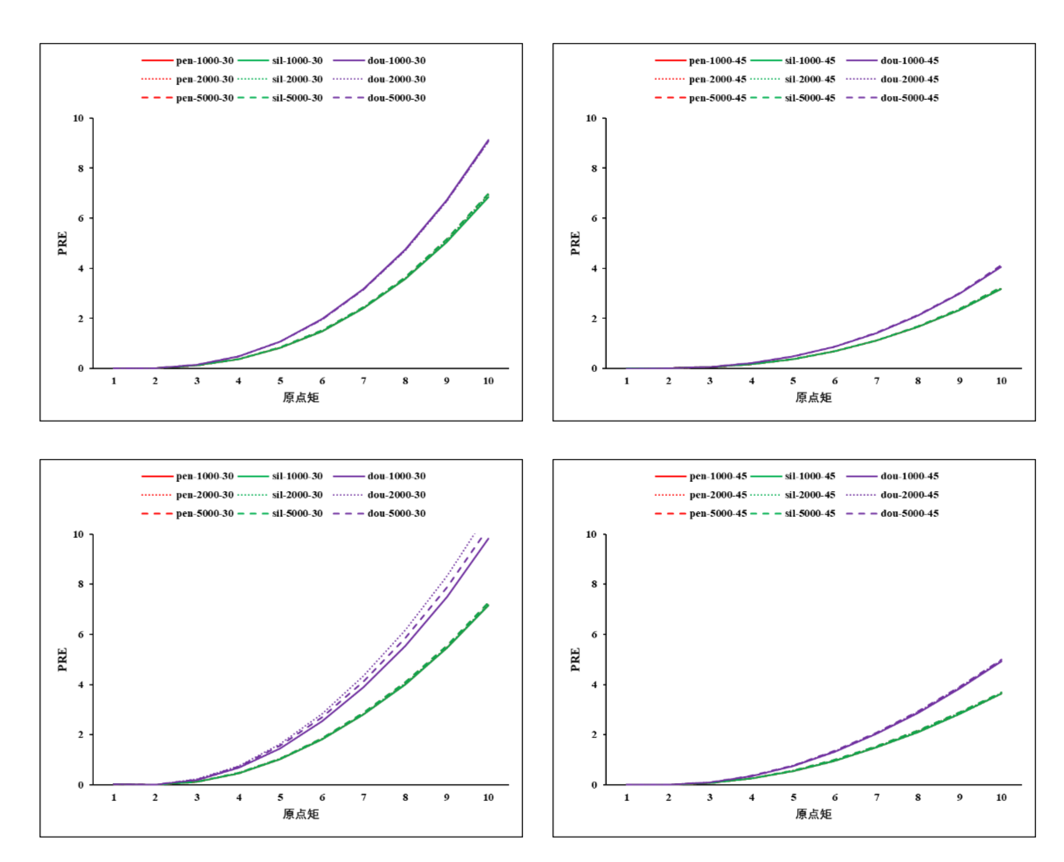


图 4　PRE（NEAT）

注：上图与下图分别为构造法和 IRT 法；Pen、Sil 与 Dou 分别代表惩罚法、Silverman 经验准则、重复平滑法；1000、2000、5000 与 30、45 分别代表考生样本量和题目量；除左下图外，其他三种带宽选择方法结果基本重合。

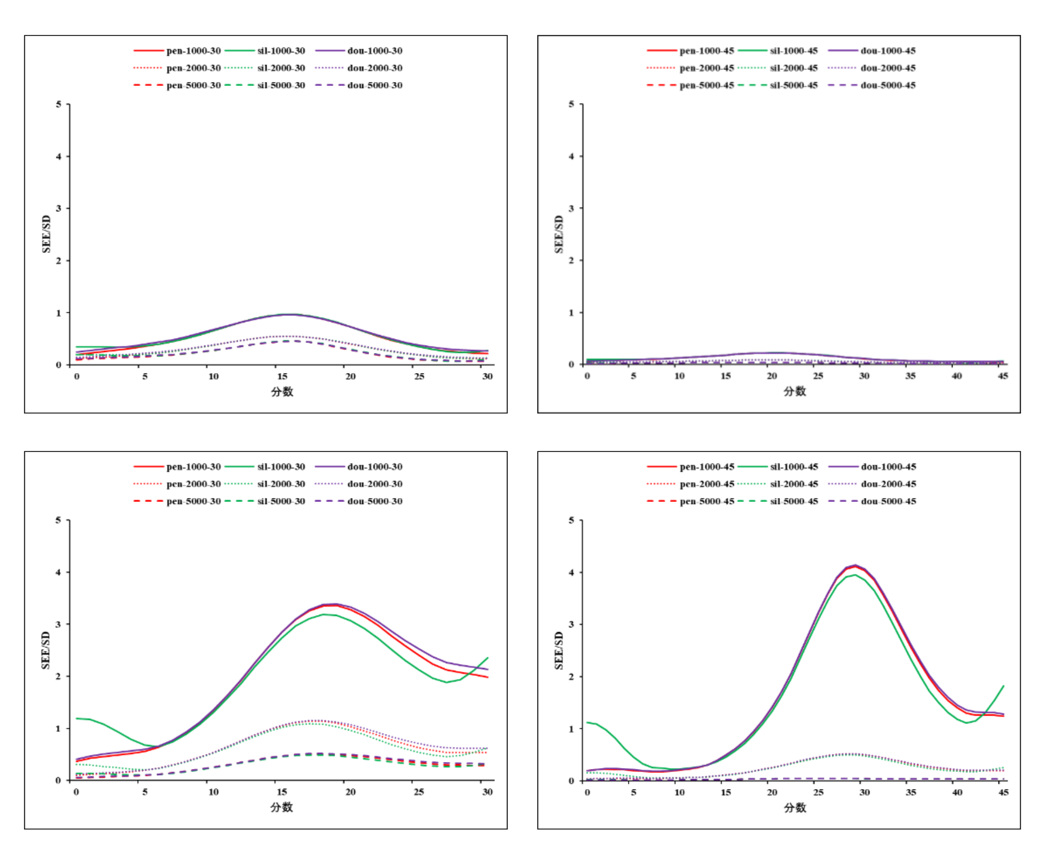


图 5　SEE（NEAT）

注：上图与下图分别为构造法和 IRT 法；Pen、Sil 与 Dou 分别代表惩罚法、Silverman 经验准则、重复平滑法；1000、2000、5000 与 30、45 分别代表考生样本量和题目量；纵轴将所有结果置于标准差单位量尺。

复平滑法。区分 EG 与 NEAT 的关键因素之一是等值群体能力差异（Kolen & Brennan, 2014）。为充分体现非等组，依据 Form2 成绩，构造高、低能力的考生群体，其平均分数差异为 18.21 分（满分 80 分）。故而，考生能力水平差异使 EG 与 NEAT 结果有别，且在同等条件下，前者误差较后者小（Dorans et al., 2008; Sinharay & Holland, 2010; Wang et al., 2008），这在本研究两大类指标中均有所体现。综合两种等值设计，惩罚法与 Silverman 经验准则表现较优。保留两位小数，二者 PRE 值相同。这可能是 Form1 分数特征与标准正态分布相近，而 Silverman 经验准则正是基于正态分布的简单算法（Wallin et al., 2017）。Andersson 和 von Davier（2014）也有相似发现，当分数分布较为平滑时，惩罚法第一部分表现优异；反之，Silverman 经验准则较好。Silverman 经验准则计算简单、直接；但当分数分布非正态时，误差可能较大。可见，各方法具有一定程度的稳健性。

5.2 考生样本量与题目数量

其他条件不变，增大样本量可降低 SEE（Kolen & Brennan, 2014），这在 NEAT 中最明显。二者间存在边际递减关系——SEE 降低的幅度随样本量增大而逐渐变小。Godfrey（2007）发现，当样本量较小时，KE 与常用等值方法间存在明显差异；但随其增大，各方法趋于一致。Kim（2014）、Liang 和 von Davier（2014）的结论类似。这是因为，等值系统误差主要来源于估计准确性、统计假设、等值设计与组间差异，受随样本量影响较小。而主要来源于抽样代表性的随机误差则会随样本量增大而减小（Kolen & Brennan, 2014）。因此，适当增加样本量可提高等值准确性。受限于数据源的题目量，本研究未探讨测验题目量较大的情况。

不计其他因素，测验题目量与信度成正比，可为等值提供有利保障；但题目增多也使分数区间扩张，分配到各分数点的考生量减少，导致误差扩大（Wang et al., 2008）。在实际情况中，随题目数量增多，信度影响等值的增量变小，而各分数的频率却会急遽减小。EG 中的 SEE 指标也体现此点。例如，Norman Dvorak（2009）发现，当题目量从 25 道增加到 75 道时，测验信度不断增加，而 KE 的均方根差异等误差指标也在随之增大。

但在 NEAT 中，锚测验题目增多可降低等值误差。锚测验是分离考生能力差异与试卷难度差异的关键因素，适当增加其长度，可更好地区分两种差异。例如，Andersson（2016）和 Kim（2014）发现，增加锚题比例可有效减小等值系统误差。锚测验长度对等值准确性的影响也受边际递减效应制约，比如 APRE 和 ASEE 的表现。除此之外，在图 5 中，当采用 IRT 法且样本量为 1000 时，在 30 分左右，题目数量增加反而导致 SEE 变大，这与 ASEE 结果不一致。对比两种条件的原始分数分布（未呈现），题目数量为 30 时，分数集中于 15 左右；题目数量为 45 时，分数集中于 25 左右。SEE 与 ASEE 结果不一致主要由于后者为考虑分数分布的加权指标（整体表现）。而仅在相同分数位置比较（局部表现）时，题目数量增加可降低 SEE，这一结论在大多数情况下（除 30 分左右外）

仍成立。未来可详细探讨锚题比例（锚题数量）对 IRTKE 的影响。

5.3 数据模拟方式

构造法的初衷是对数据模拟方式偏向性的质疑。结果表明，基于构造法的指标较 IRT 法小，其中以 NEAT 的 ASEE 最为突出。构造法与 IRT 法的 SEE 形状存在较大差异，但二者与其他因素间不存在明显的交互作用。结合表 3 与表 4 推测，相较于 IRT 法，构造法获得的分数偏高且更为集中，故而低分段人数较少，从而导致 SEE 偏大。von Davier 等人（2004）详细描述并解释过 KE 中这种常见的“两端高中间低”的情况。但该差异是否由数据模拟方式造成，能否影响分数解释，尚无定论。须慎重对待仅采用 IRT 法的研究结果，可选择较为中立的构造法开展研究，未来仍需深入探索二者间的差异与原因。

6 结论

IRTKE 具备 IRTOSE 和 KE 的优异特性，本研究从不同等值设计和数据模拟方式角度，探讨带宽选择方法、样本量、题目数量对其的影响，主要得出以下结论：（1）在 EG 中，惩罚法、Silverman 经验准则和重复平滑法表现相似。（2）在 NEAT 中，惩罚法与 Silverman 经验准则表现较佳。特别地，在极端分数附近，Silverman 经验准则略逊于其余两种方法；而在中间偏高分数处，表现较优。（3）在一般情况下，增大考生样本量和题目数量均可降低等值误差，该现象在 NEAT 中最为明显。（4）数据模拟方式可干扰结果，在涉及方法比较的研究中，尤其是当待比较的方法与数据模拟方式共享某种理论时，应尽量避免其对结果的误导。

致谢：感谢华南师范大学心理学院研究生甄锋泉在模拟代码编写过程中提供的帮助！

参考文献

罗莲. (2008). 一种新的等值方法: 核等值法. *心理学探新, 28*(2), 69–74.

王少杰, 张敏强, 李拓宇, 梁正妍. (2020). 核等值: 一种观察分数等值体系. *心理科学进展, 28*(5), 855–870.

Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement, 53*(4), 459–477.

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software, 55*(6), 1–25.

Andersson, B., & von Davier, A. A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement, 51*(3), 223–238.

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika, 82*(1), 48–66.

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software, 68*(7), 1–22.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

De Ayala, R. J., Smith, B., & Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Applied Psychological Measurement, 42*(2), 155–168.

Dorans, N. J., Liu, J. H., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81–97.

Godfrey, K. E. (2007). *A comparison of kernel equating and IRT true score equating methods* (Unpublished doctorial dissertation). The University of North Carolina at Greensboro.

González, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Springer.

Hagge, S. L., & Kolen, M. J. (2012). Effects of group differences on equating using operational and pseudo-tests. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2)* (pp. 45–86). CASMA, The University of Iowa.

Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed - score kernel equating. *Journal of Educational Measurement, 51*(2), 201–211.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Jiang, Y. L., von Davier, A. A., & Chen, H. W. (2012). Evaluating equating results: Percent relative error for chained kernel equating. *Journal of Educational Measurement, 49*(1), 39–58.

Kim, H. Y. (2014). *A comparison of smoothing methods for the common item nonequivalent groups design* (Unpublished doctorial dissertation). University of Iowa.

Kim, S., & Lu, R. (2018). The pseudo-equivalent groups approach as an alternative to common-item equating. *ETS Research Report Series, 2018*(1), 1–13.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Leôncio, W., & Wiberg, M. (2017). Evaluating equating transformations from different frameworks. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 101–110). Springer.

Liang, T., & von Davier, A. A. (2014). Cross-validation: An alternative bandwidth-selection method in kernel equating. *Applied Psychological Measurement, 38*(4), 281–295.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings”. *Applied Psychological Measurement, 8*(4), 453–461.

Norman Dvorak, R. L. (2009). *A comparison of kernel equating to the test characteristic curve method* (Unpublished doctorial dissertation). University of Nebraska.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). *A test of the adequacy of linear score equating models*. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). Academic Press Inc.

Powers, S. J., & Kolen, M. J. (2012). Using matched samples equating methods to improve equating accuracy. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2)* (pp. 87–114). CASMA, The University of Iowa.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica, 77*(4), 329–352.

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*(3), 261–285.

von Davier, A. A., & Chen, H. W. (2013). The kernel Levine equipercentile observed-score equating function. *ETS Research Report Series, 2013*(2), i–27.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.

Wallin, G., Häggström, J., & Wiberg, M. (2017). How to select the bandwidth in kernel equating-An evaluation of five different methods. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 91–100). Springer.

Wang, S. J., Zhang, M. Q., & You, S. (2020). A comparison of IRT observed score kernel equating and several equating methods. *Frontiers in Psychology, 11*, Article 308. https://doi.org/10.3389/fpsyg.2020.00308

Wang, T. Y., Lee, W. C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632–651.

Wiberg, M. (2016a). Alternative linear item response theory observed-score equating methods. *Applied Psychological Measurement, 40*(3), 180–199.

Wiberg, M. (2016b). Ensuring test quality over time by monitoring the equating transformations. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 239–251). Springer.

Effects of Several Factors on IRT Observed Score Kernel Equating

Wang Shaojie, Zhang Minqiang, Huang Feifei, Huang Lifang, Yuan Qiting

(School of Psychology, South China Normal University, Guangzhou, 510631)

Abstract

Attributing to its advantages of pre-smoothing and continuization of score distributions, kernel equating has been testified and shown equivalent to or better than other equating methods, especially traditional ones, in the aspect of equating accuracy and stability. IRT observed score kernel equating is formed by integrating kernel equating and IRT observed score equating. Few researches have focused on evaluating its performance systematically. Therefore, bandwidth selection method, sample size, test length, equating design, and data simulation methods were investigated about their influence on it.

To ensure ecological validity, data from a large-scale assessment were used as the sampling pool. IRT data simulation method and pseudo tests and pseudo groups simulation method were used to avoid the simulation preference in random Equivalent Groups design (EG) and Non-Equivalent groups with Anchor Test design (NEAT). In detail, bandwidth selection methods included Penalty method, Silverman's rule of thumb method, and Double smoothing method. Levels of sample size were 1000, 2000, and 5000. Meanwhile, test containing 30 items and 45 items were considered. Finally, local criteria and universal criteria were computed, the former of which were Percent Relative Error (PRE) and Standard Error of Equating (SEE), and the latter of which were Averaged Percent Relative Error (APRE) and Averaged Standard Error of Equating (ASEE).

It was found out that in EG, regarding local criteria, PRE increased as central moment became higher, which also meant that the distribution difference before and after equating was enlarged. Nonetheless, considering that PRE was formed by multiplying initial difference with 100, bandwidth selection methods performed alike. On the other hand, PRE was significantly reduced by increasing sample size and lengthening tests, especially by the latter one. Similar to PRE, when it came to SEE, there was no difference between effect of bandwidth selection methods. Larger sample size rendered less random error, which was contrary to test length. Furthermore, curves of SEE were "high at left but low at right" for pseudo tests and pseudo groups method, and "low at left but high at right" for IRT simulation method. As for universal criteria, APRE among bandwidth selection methods were alike, which were all small. Effects of sample size and test length were same as observed in local criteria. There was no significant difference between ASEE for two data simulation methods.

In NEAT, regarding local criteria, PRE increased as central moment became higher. The results of Penalty method and Silverman's rule of thumb method coincided, which were superior to others. And this trend was more evident when test is shorter. PRE was significantly reduced by lengthening tests as in EG, but not by increasing sample size. To be mentioned was the results that PRE for Double smoothing method was most influenced by sample size when test included 30 items and IRT simulation method was used, which indicated some interactions among them. When it came to SEE, bandwidth selection methods performed alike, only showing discrepancies at extreme scores. Increasing sample size and lengthening test could reduce random error. Meanwhile, distribution of SEE for pseudo tests and pseudo groups method was more stable than that for IRT method. As for universal criteria, the trends for APRE and ASEE were same as those in local criteria.

To summarize, performances of bandwidth selection methods were similar in EG, but Penalty method and Silverman's rule of thumb method prevailed in NEAT. Bandwidth selection, sample size, and test length affected IRT observed score equating together. Preference of data simulation methods was spotted, which suggested researchers that multiple simulation methods and designs should be conducted before final conclusions are drawn in the field of comparison of equating method. Further study should focus more on the systematic evaluation of equating.

Key words IRT observed score kernel equating; bandwidth selection methods; equating design; data simulation methods